Application of Artificial Intelligence for Chemical Inference. XXV. A Computer Program for Automated Empirical ^{13}C NMR Rule Formation

By

Tom M. Mitchell and Gretchen M. Schwenzer*

Contribution from the Department of Computer Science,
Stanford University,
Stanford, California 94305

(February 1977)

Abstract: A computer program which generates empirical rules associating ^{13}C NMR shifts with local structural environments is described. The program uses a heuristic method to search for common structural features for those carbon atoms exhibiting similar shifts. Rules have been generated by our program from a combined set of acyclic amine and paraffin data. Examples of these rules are presented, and their performance as a tool for structure elucidation is examined.

Introduction

Recent computer studies[1,2,3,4] have explored the automated application of ^{13}C NMR techniques to structure elucidation problems. The increasing popularity of ^{13}C techniques and the increasing bulk of available data have motivated us to develop a computer program which generates empirical ^{13}C NMR rules.

A natural abundance ^{13}C NMR spectrum which is fully proton-decoupled consists of a number of sharp peaks corresponding to the resonance frequencies in the applied magnetic field of the various types of carbon atoms present in the sample. A ^{13}C shift is the amount a peak position deviates from a reference peak of a standard compound usually tetramethylsilane (TMS). Methods for obtaining empirical rules which correlate ^{13}C shifts with local structural environments within a class of compounds are cited in the literature.[5,6,7,8] In the traditional parameter set approach to rule formation the chemist sorts through a large amount of data by hand, searching for structural features which appear to correlate with ^{13}C shifts. The total ^{13}C shift is then given as a function of these structural features. The functional form chosen is usually one which presumes a linearly independent relationship between the structural features. A curve fitting procedure is used to obtain optimal settings for parameters in the equation. The selection of pertinent structural features and the selection of a functional form are both difficult decisions for which it may be easy to overlook the correct choice.

We have written a computer program[9] which generates empirical rules associating a ^{13}C NMR shift with the local structural environment of a carbon atom. The chemist must provide the program with a training set of known structures and their assigned ^{13}C spectra. In addition he must select a language of discrete valued atom features (i.e., "atom type", "number of neighbors", etc.) in which the program is to express substructures. The program then builds chemical substructures which characterize the environment of carbon atoms exhibiting similar ^{13}C shifts.

A set of rules in the following format is generated:

(substructure description) $\xrightarrow{\text{implies}}$ (range of characteristic ^{13}C shift)

If the substructure to the left of the arrow is present within some molecule then for a specified atom within the substructure a ^{13}C shift in the range given to the right of the arrow will be observed. For example, the following rule was generated by the program when presented with a training set of combined paraffin and acyclic amines.

$$CH_3^{*}\text{-}CH_2\text{-}CH_2\text{-}CH_2\text{-}X \rightarrow 14.0 \text{ ppm} \leq \delta_C \leq 14.7 \text{ ppm}$$

The asterisk in the substructure description denotes the atom for which the shift is predicted. The X stands for any non-hydrogen atom. For the above rule δ_C is given in ppm downfield from TMS.

A second program is capable of applying the rules to molecules to predict and assign their ^{13}C NMR spectrum. Also, given an unknown ^{13}C NMR spectrum and a list of candidate molecules, it can predict a spectrum for each candidate and rank the candidates by the similarity of their predicted spectra to the unknown spectrum. Although it is possible to apply the rules to any molecule, the value of the prediction depends upon the similarity of the molecule to the training set which was used to generate the rules.

A reliable automated method of generating and applying ^{13}C NMR rules under constraints supplied by the chemist can be a valuable tool in increasing the efficiency of analyzing ^{13}C NMR data and deducing structural information from the ^{13}C NMR spectra of unknowns.

Method

A. **Rule Generation.** The algorithm employed to generate the ^{13}C NMR rules closely parallels the algorithm in the Meta-DENDRAL[10,11] program which generates empirical rules of molecular fragmentation from mass spectral data. Both programs represent molecules in a chemical graph notation where the atoms are represented as nodes and bonds are represented as arcs between nodes. A heuristic method[12] is used to search through a space of possible substructures. For ^{13}C NMR rule formation this search is directed toward finding substructure descriptions of the characteristic local environment about carbons exhibiting similar ^{13}C shifts.

The chemist is given some control over the character of the generated rules. He must select the language of atom features used to generate substructures for the rules. In this work a topological representation of structures was employed which included "atom type" (C,N,etc.) and "number of non-hydrogen neighbors" as atom features (hydrogens are not represented as distinct atoms in this representation). Although stereochemical terms were not required to analyse the amines and paraffins, the addition of stereochemical terms as atom features will be a topic of future work. The chemist must also assign values to two parameters which control the specificity/generality of the generated rules. The parameter MINIMUM-EXAMPLES specifies the minimum number of data points which a rule must explain within the training set. A second parameter, MAXIMUM-RANGE, specifies the maximum allowed width of a rule's predicted shift range. A chemist who wants to find only the most common general trends in a set of data may set these two parameters to obtain a small number of fairly general rules with moderate sized prediction ranges. When interested in a more exacting analysis, he may set the parameters to obtain a larger number of rules containing relatively specific substructures and correspondingly smaller prediction ranges.

Rule generation may be viewed as a search which starts with the very general seed rule $C \rightarrow -\infty<\delta_C<\infty$ (where C may be any carbon atom with any atom properties and δ_C is the observed shift), and proceeds to expand upon this rule by adding new atoms and atom features to the rule substructure in an attempt to narrow the predicted range of shifts. Since new atoms and atom features may be added to a substructure in many ways, each rule considered may be expanded into many offspring. The first few steps of the rule search in generating the results discussed below are shown in Figure 1.

At each step in the search a single atomic feature from the user selected list is added to all the atoms a given number of bonds away from the central carbon. In Figure 1 we see that the seed rule is expanded by considering all possible values for "number of neighbors" of the central carbon. In this example only "atom type" and "number of neighbors" are allowed as atom features (note that the feature "atom type" was already specified as carbon in the seed rule). Each resulting level 1 substructure is in turn expanded in level 2 by adding either an "atom type" or "number of neighbors" specification to each atom one bond away from the central carbon. The search is then continued along each of these new paths. At each step each newly generated substructure is associated with a range of ^{13}C shifts. This range is determined by searching for occurances of the substructure within the training set molecules. When the substructure is found within a molecule in the training set, the observed shift for the associated carbon is recorded. The shift range associated with the substructure is simply the range of all such observed shifts.

Each step taken in the rule search is evaluated in terms of the shift range of the resulting substructure. If the shift range is found to be narrower than the range of the parent rule, then the added specification is considered to be useful, and the search is continued from the new version. If an added specification does not narrow the shift range, that path of the search is aborted. More specifically for each newly generated potential rule, the following action is taken:

```
IF the shift range of the new version is not narrower than the shift range of its parent,
    THEN abort this branch of the search;
ELSE IF the number of applications of the new version in the training data is less than some predefined value, MINIMUM-EXAMPLES,
    THEN add its parent to the list of rules and do not consider the new version (the user may set a control flag which prevents the parent from becoming a rule in this case);
ELSE IF the shift range of the new version is smaller than some predefined value, MAXIMUM-RANGE,
    THEN add the new version to the list of rules;
ELSE consider the versions generated by adding a new level of feature specifications.
```

The program runs until all branches of the search have been explored. At that point a set of rules will have been generated which covers all the data points in the training set. Since rules with slightly different substructures may have been generated which cover the same data in the training set, the rule set may be redundant. The set of rules is therefore passed to a routine which selects a less redundant subset of the rules. The algorithm for selecting a subset of rules is summarized below:

1. Assign to each rule the score peaks/wrange^2 where peaks is the number of data peaks covered by the rule in the training data, and wrange is the width of the shift range for the rule.

2. Select the rule with the highest score.

3. Delete the data points covered by the newly selected rule, and reevaluate the scores of all remaining rules.

   If a rule covers no data points, remove it.

4. If there are more rules left, go to (1).

The intent of the algorithm is to select during each iteration the strongest rule, then to weaken rules with evidence which overlaps with it. The effect of this procedure is to select a subset of the strongest rules covering the same set of data as the entire set of rules.

B. Use of rules for structure selection. In order to test the utility of the rules as a tool for structure elucidation a second program was written which uses the rules to select from a list of candidate structures the one whose predicted spectrum most closely matches the given unknown spectrum. The structure selection test involves two steps: 1)predict the spectra of a set of candidate molecules using the rules, and 2)compare the predicted spectra with the unknown spectrum.

1. Predicting a spectrum. Rules are applied to a molecule simply by searching for places where the rule substructure fits into the molecule. A graph matching routine is used to find a mapping of atoms in the rule substructure to atoms in the molecule. When a match is found the shift range associated with the rule is predicted for the associated carbon atom. An example of predicting a spectrum is shown in Figure 2. For each carbon in 4,4-Dimethylheptane the rules which apply to it are shown. For each rule shown its substructure and atom features map into 4,4-Dimethylheptane.

Often several rules apply to predict ranges for the same carbon atom. If the predicted ranges are consistent with each other (i.e., one of the predicted ranges is contained in the others), then the narrowest predicted range is used. This is illustrated by the rules which explain C_2 and C_5 in Figure 2. Since one of the three predicted shift ranges is contained within the others, it is selected. The rationale for this decision is that the actual shift should fall into all of the predicted shift ranges; therefore assume the most stringent prediction. If the predicted ranges overlap incompletely or are disjoint (this can occur for structures outside the training set), then the ranges which are inconsistent with the narrowest range are merged to arrive at a final predicted range for the carbon atom.

2. Comparing Predicted and Observed Spectra. Spectrum assignment is accomplished by assigning each atom's predicted range to the closest observed shift in the unknown spectrum. In order to be a valid assignment, the condition

$$\begin{pmatrix}\text{no. of carbons}\\ \text{in structure}\end{pmatrix} - \begin{pmatrix}\text{no. of observed}\\ \text{shifts}\end{pmatrix} \geqslant \begin{pmatrix}\text{no. of multiply}\\ \text{assigned observed shifts}\end{pmatrix}$$

must be satisfied. If the assignment satisfies this constraint it is kept. Otherwise, the required number of multiply assigned observed shifts are reassigned. The reassigned atoms and shifts are selected by approximating the negative contribution to the comparision score of reassigning each multiply assigned atom. The reassignments corresponding to the greatest comparison score are selected.

Once the most feasible assignment is found, the predicted shift of each carbon is compared with the assigned shift in the unknown and a comparison score for the molecule is obtained. The comparison score reflects the degree of closeness of a structure's predicted spectrum to the unknown spectrum. This score is:

$$\text{Comparison score} = \sum_{\substack{\text{predicted}\\ \text{ranges}}} \frac{1}{\text{hrange}} \cdot \left(1 - \frac{|\text{oshift} - \text{pshift}|}{\text{hrange}} \right)$$

where

hrange = half the width of the predicted range
pshift = the midpoint of the predicted range; the predicted shift
oshift = the shift of the assigned peak in the unknown spectrum; the observed shift.

Since this function is inversely proportional to the width of the predicted range, narrow predicted ranges carry more weight than wide predicted ranges. The comparison score is positive if the observed shift is within the predicted range, and negative otherwise.

The structure selection program contains a user set parameter EXPECTED-ERROR to account for the instrument resolution, inaccuracies in the rules, and other sources of error in the predicted or unknown spectra. Each predicted range in each predicted spectrum is enlarged by the value of the parameter setting. In Figure 2 the predicted ranges have had 2 ppm added to both their upper and lower bounds.

Results

In order to test the viability of our approach to automated ^{13}C NMR rule formation, we selected paraffins and acyclic amines for test compounds since they are known to exhibit a strong correlation between ^{13}C shift and local structure.[5,6,7,8]

A. Rules. Rules were generated from a combined set of 22 paraffins and 47 acyclic amines selected as a representative subset of those cited in the literature.[6,7] Structures with empirical formula C_9H_{20} or $C_6H_{15}N$ were excluded from the training set since these were to be used later as unknowns for a structure elucidation test.

The shifts for the paraffins were incremented by 0.6 ppm in order to account for the consistent difference between the two sets of data. This compensates for the different solvents used for the amines (benzene) and paraffins (dioxane). The program was given the atom features "atom type" (i.e., C,N,etc.) and "number of neighbors" with which to construct substructures. The parameter settings were MINIMUM-EXAMPLES = 2 and MAXIMUM-SHIFT-RANGE = 2.0 ppm. The program generated a set of 138 rules which covered all of the 435 data peaks. Examples of the rules obtained are given in Figure 3.[13] Of the rules generated 41 had a predicted shift range greater than 2.0 ppm since they covered substructures which did not occur often enough in the data to allow generating a rule with 2 instances and a narrow shift range. As might be expected, the rules with wide shift ranges had general substructures which applied in many places such as Rule 4 in Figure 3 while the rules with narrower shift ranges carried more detailed substructures applying in fewer places such as Rule 3 in Figure 3. Often multiple rules were found to cover the same data point. For example Rule 4 in Figure 3 applies wherever Rule 3 applies.

Since the program is designed to build substructures by including only atoms and atom features which improve a rule's performance in the training data, some rule substructures may omit specifications of atoms and features which could be added without affecting the evidence of a rule. Rule 1 in Figure 3 says that whatever the type of the atom next to the methyl group, the shift will be (14.32 to 14.70). Actually, this rule was formed from situations in which atom 2 was always carbon. However, since within the training data the addition of the atom type could not change the rule's evidence, the type of atom 2 remained unspecified. The program did consider including the atom type, but since it did not change the rule's evidence the type specification was not included. If the training set had included situations in which the specification of the atom type would have improved the prediction, the specification would have been added.

The example given above illustrates a design decision which was built into the program. For a given rule there may be several possible substructure descriptions which perform equally well within the training set data. The program selects one of these substructures as the left hand side of the rule. A tradeoff exists between selecting a specific substructure with many atoms and features or a more general substructure with a relatively general description of the local environment. The more specific substructure is less likely to apply in subsequent molecules. A less specific substructure is more likely to apply, but also may apply in cases where it should not (as is the case for Rule 1 of Figure 3). We are working on an extension to the algorithm which will allow the program to reason more thoroughly with this issue, and to refine its decision when presented with additional data.

Other properties of the rules are illustrated by the examples in Figure 3. In Rule 1 the neighbor property for atom 4 means that atom 4 is connected to three atoms one of them being atom 3 and two other atoms which were not given in the substructure. Rule 2 gives the shift of a quartenary carbon. Each carbon adjacent to the quaternary carbon may be connected to additional atoms outside the substructure. Rules 3 and 4 illustrate the dependence of the ^{13}C shift range upon the number of nearest (alpha), next nearest (beta), and further levels of neighboring carbon atoms. This is similar to the result that Lindeman and Adams obtained.[6] Rule 4 says that atom 3 has three alpha carbons and any number of beta carbons and gives a broad prediction range. Rule 3 refines this prediction range giving a rule for the case of three alpha carbons and two beta carbons. Rules 5 and 6 are both for a tertiary amine and illustrate the influence of an additional beta carbon on the tertiary carbon.

The rules correctly covered all the data in the training set with an average prediction range of 2.7 ppm. Notice that the average prediction range could have been made smaller (with more resulting rules) or larger (with fewer rules) by adjusting the parameters MAXIMUM-RANGE and MINIMUM-EXAMPLES.

B. Structure selection of acyclic amines and paraffins. The structure selection program was supplied an exhaustive list of C_9H_{20} structural isomers (35 distinct structures[14]) which had been omitted from the training set data used in the rule formation step. The program used the set of rules to predict a spectrum for each structural isomer. An unknown spectrum belonging to one of the C_9H_{20} structures was supplied and the program was asked to rank the candidate structures by comparing each predicted spectrum with the unknown spectrum. The number of carbons that correspond to an observed peak was not included with the unknown spectrum. The test was repeated with each of the 24 C_9H_{20} spectra from Lindeman and Adams[6]. The results of this test and another test involving $C_6H_{15}N$ amines, are given in Table 1. Notice that the program ranked the correct structure either first or second for 23 of the 24 C_9H_{20} spectra and for 10 of the 11 $C_6H_{15}N$ spectra. The value of the parameter EXPECTED-ERROR in the spectrum comparison which allows the user to widen the predictive range of the rules was varied. The program performed best on this data and with this set of rules with EXPECTED-ERROR equal to 2 ppm. *It is difficult to sort out the reasons* for this setting being optimal. Two relevant factors are the peak comparison function (the value decays to zero when the observed shift is at the edge of the predicted range) and the fact that the rules were trained on different molecules than used here.

We examined in detail the program's analysis of the 2,2,3,3 - Tetramethylpentane spectrum. In this case the program ranked the correct structure ninth out of the 35 candidates. We found a combination of effects which compounded to give the low ranking to the correct structure. These effects were also apparent to a lesser degree in the analysis of other spectra, and point to possible improvements to our approach.

Some predicted shift ranges for the correct structure were quite wide. For 2,2,3,3 - Tetramethylpentane, seven of the nine predicted ranges were wider than 6 ppm due to two factors. First, wide predicted shifts may arise for carbons with specific local environments which were not found in the training data. In this case only a very general rule (if any) might be available to predict a range for the carbon. Second, a wide predicted shift may also arise when two rules with disjoint predicted ranges apply to the same carbon. In this case, at least one of the rules must have been incorrect. Either its prediction range should have been wider or the substructure made more specific. This problem is related to the tradeoff discussed earlier between using general or specific versions of the structure.

Since the predicted ranges are assigned to the rules only on the basis of those examples observed in the training set it is possible for shift ranges to be too narrow. In this case, the inadvertently narrow predicted range may incorrectly penalize the correct candidate structure.

Peak intensity information which gives the number of carbon atoms corresponding to an observed peak was not used. This information could be used by the program as a constraint on assignment of predicted ranges to peaks in the unknown spectrum. The expected effect would be lower comparison scores for incorrect structures without significant change to the comparison score of the correct structure.

Conclusion

The rules generated by the program are of the form substructure implies shift range. This form differs in several respects from the traditional parameter set approach of weighting and summing predefined structural features. (1) Each rule has a distinct predicted shift range. Thus the ruleset can include rules of varying detail. The usual expected error for the parameter set approach is a single number - the standard deviation of the training set from the fitted curve. The prediction of a shift range rather than a single number is an advantage when analyzing carbon atoms that exhibit magnetic nonequivalence. Magnetic nonequivalence results in different chemical shifts for two identical groups in molecules having an asymmetric carbon atom. The parameter set approach which attempts to predict the arithmetic mean of the observed shifts will always be in error. (2) The rule format used within the program is ideally suited for "reading backwards", that is, the appearance of a peak in an unknown spectrum implies a structural feature. We hope to make use of this ability in a future program which will use the rules to infer likely structures from an unknown spectrum. (3) Whereas the parameter set approach attempts to break down the total shift in terms of contributions from the selected structural features, our rules predict only the total ^{13}C shift characteristic of the substructure. This is a major difference, and there is both an advantage and a disadvantage in our approach. The disadvantage is that by overlooking contributions to total shift our program is forced to generate a new rule to cover every distinct ^{13}C shift. Although a large set of rules is difficult for people to grasp quickly, computer programs may quite easily apply the rules. Predicting a ^{13}C spectrum required approximately one minute of CPU time for the average molecule in this study. The advantage of predicting total shifts over hypothesizing partial contributions is in avoiding initial biases as to what contributes to the shift and how these contributions are to be combined. In the parameter set approach preselected structural features are combined in a predefined functional form. Our approach bypasses the bias of an assumed functional form, and introduces only a weak bias concerning which structural features may be considered. The program may consider any structural feature which can be expressed within the language of atom features selected by the chemist.

The performance of the rules in discriminating among similar structures not included in the training set data demonstrates the general content of the rules. Although the procedure of predicting spectra for all possible structural isomers, then comparing against the unknown spectrum is an inefficient approach to structure elucidation (the number of possible structural isomers increases rapidly with the size of the molecule), it is nevertheless a valid test of the information content of the rule set.

Acknowledgements. This work was supported by the National Institutes of Health under grant 5R24 00612-07 and by the Advanced Research Projects Agency under grant DAHC 15-73-C-0435.

Computer resources were provided by the SUMEX facility at Stanford University under National Institutes of Health grant RR-00785.

We are grateful to many people with whom we have discussed and planned this work. We wish to thank Bruce G. Buchanan for helping adapt the Meta-DENDRAL algorithm to the ^{13}C NMR domain and for repeated assistance and support throughout this work. Carl Djerassi originally suggested ^{13}C NMR as a possible domain for rule generation. Dennis Smith, Raymond Carhart, Tomas Varkony, and Jim McDonald have all provided useful suggestions concerning the form of the rules and the algorithm for their generation. We also express our thanks to William C. White for providing parts of the Meta-DENDRAL program code for use in this work.

References

(1) W. Bremser, M. Klier and E. Meyer, Org. Magn. Resonance, 7, 97 (1975).

(2) Raymond E. Carhart and Carl Djerassi, J. Chem. Soc., Perkin Trans. 2, 1753 (1973).

(3) Barbara A. Jezl and David L. Dalrymple, Anal. Chem., 47, 203 (1975).

(4) R.S. Schwarzenbach, J. Meili, H. Könitzer and J.T. Clerc, Org. Magn. Resonance, 8, 11 (1976).

(5) D.M. Grant and E.G. Paul, J. Am. Chem. Soc., 86, 2984 (1964).

(6) L.P. Lindeman and J.Q. Adams, Anal. Chem., 43, 1245 (1971).

(7) Hanne Eggert and Carl Djerassi, J. Am. Chem. Soc., 95, 3710 (1973).

(8) Joseph E. Sarneshi, Henry L. Surpreant, Frederick K. Molen and Charles N. Reilley, Anal. Chem., 47, 2116 (1975).

(9) The programs are written in INTERLISP and run on the DEC10 SUMEX-AIM computer resource at Stanford University.

(10) B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg and Carl Djerassi, J. Am. Chem. Soc., 98, 6168 (1976).

(11) B.G. Buchanan, Proceedings of the NATO Advanced Study Institute on Computer Oriented Learning Processes, Bonas, France, 1974.

(12) A heuristic method uses information about the nature of a problem to guide the search for a solution. Our program relies upon (among other things) information supplied by the user

about the type of rules which are of interest.

(13) A conplete listing of the rules generated by the program is available from the authors.

(14) Raymond E. Carhart, Dennis H. Smith, Harold Brown and Carl Djerassi, J. Am. Chem. Soc., 97, 5755 (1975).

Figure Captions

Figure 1. Partial schematic diagram of the rule search, Rule substructures are expanded by adding new substructure specifications. The observed shift range in the training data directs the search. δ_C values are approximate and are given in ppm downfield from TMS. The '*' identifies the central carbon to which the shift is assigned. 'X' indicates that any non-hydrogen "atom type" is allowed.

Figure 2. Application of rules to 4,4-Dimethylheptane for spectrum prediction. $\delta_C(n)$ is the shift observed for the atom n in ppm downfield from TMS.

Figure 3. Sample rules generated by the program. δ_C is given in ppm downfield from TMS.

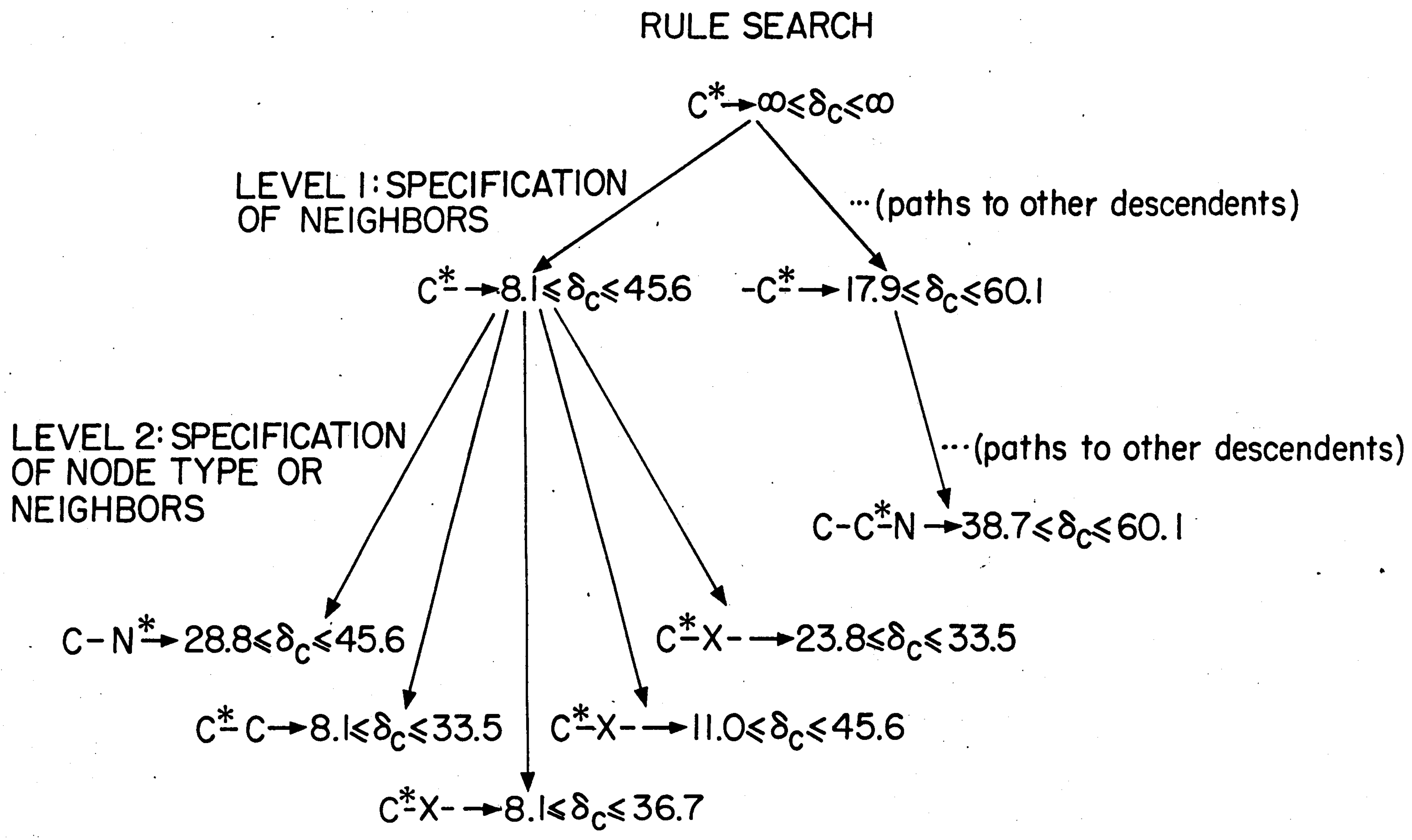
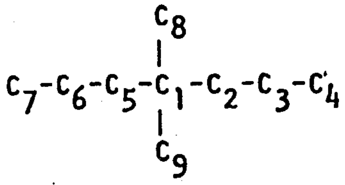

RULE SEARCH
$C^* \rightarrow \infty \leq \delta_C \leq \infty$
LEVEL 1: SPECIFICATION OF NEIGHBORS
···(paths to other descendents)
$C^*- \rightarrow 8.1 \leq \delta_C \leq 45.6$
$-C^*- \rightarrow 17.9 \leq \delta_C \leq 60.1$
LEVEL 2: SPECIFICATION OF NODE TYPE OR NEIGHBORS
···(paths to other descendents)
$C-C^*-N \rightarrow 38.7 \leq \delta_C \leq 60.1$
$C-N^* \rightarrow 28.8 \leq \delta_C \leq 45.6$
$C^*-X- \rightarrow 23.8 \leq \delta_C \leq 33.5$
$C^*-C \rightarrow 8.1 \leq \delta_C \leq 33.5$
$C^*-X- \rightarrow 11.0 \leq \delta_C \leq 45.6$
$C^*-X- \rightarrow 8.1 \leq \delta_C \leq 36.7$

Figure 1

4,4-Dimethylheptane



Observed Spectrum	15.5	17.9	27.6	33.4	45.4
Predicted Spectrum	(12.0 17.4)	(15.9 26.3)	(25.1 31.5)	(27.7 37.6)	(42.7 46.9)

Carbon Atom	Rules which apply to the carbon atom					
	Rule#	Substructure	Atom number	Atom type	Number of non-H neighbors	Prediction
C_2 & C_5	1	1 | 5-4-3-2-7 | 8	1,5,7,8 2 3 4	C Any C Any	≥1 4 2 2	$44.70 \leqslant \delta_C(3) \leqslant 44.90$
	2	| -4-3-2- |	2 3 4	Any C Any	4 2 2	$41.08 \leqslant \delta_C(3) \leqslant 45.13$
	3	1-2-3	1,3 2	C C	≥1 2	$17.90 \leqslant \delta_C(2) \leqslant 56.91$
C_4 & C_7	4	1-2-3-4	1 2,3 4	C C C	1 2 ≥1	$14.06 \leqslant \delta_C(1) \leqslant 15.40$
C_8 & C_9	5	7 | -2-3-4- | 8	2,4 3 7 8	Any Any C Any	2 4 1 1	$27.10 \leqslant \delta_C(7) \leqslant 29.52$
	6	3 | 8-2-1 | 7	1 2 3,7,8	C Any C	1 4 ≥1	$23.80 \leqslant \delta_C(1) \leqslant 31.86$
C_1	7	3 | 8-2-1 | 7	1,3,7,8 2	C C	≥1 4	$29.69 \leqslant \delta_C(2) \leqslant 35.60$
C_3 & C_6	8	1-2-3-4	1 2,3 4	C C C	1 2 ≥1	$17.90 \leqslant \delta_C(2) \leqslant 24.30$
	9	1-2-3	1,3 2	C C	≥1 2	$17.90 \leqslant \delta_C(2) \leqslant 56.91$

Figure 2

^{13}C NMR Rules

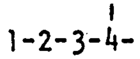
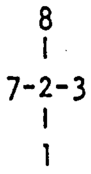
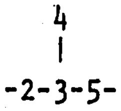
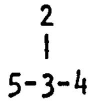
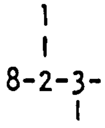
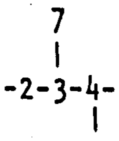
Rule#	Substructure	Atom number	Atom type	Number of non-H neighbors	Prediction	#peaks explained in training set
1		1	C	1	$14.32 \leqslant \delta_C(1) \leqslant 14.70$	4
		2	any	2		
		3	any	2		
		4	C	3		
2		1	C	≥1	$29.69 \leqslant \delta_C(2) \leqslant 35.60$	15
		2	C	4		
		3	C	≥1		
		7	C	≥1		
		8	C	≥1		
3		2	C	2	$31.67 \leqslant \delta_C(3) \leqslant 37.40$	5
		3	C	3		
		4	C	1		
		5	C	2		
4		2	C	≥1	$25.08 \leqslant \delta_C(3) \leqslant 48.20$	28
		3	C	3		
		4	C	≥1		
		5	C	≥1		
5		1	C	1	$48.09 \leqslant \delta_C(2) \leqslant 50.09$	3
		2	C	3		
		3	N	3		
		8	C	1		
6		2	C	2	$54.85 \leqslant \delta_C(3) \leqslant 56.54$	2
		3	C	3		
		4	N	3		
		7	C	1		

Table 1. Results of Structure Ranking Experiment

Identity of unknown spectrum	Ranking of correct structure out of 35 possible C_9H_{20} isomers
n - Nonane	1
2 - Methyloctane	1
3 - Methyloctane	1
4 - Methyloctane	1
2,3 - Dimethylheptane	1
2,4 - Dimethylheptane	1
2,5 - Dimethylheptane	1
2,6 - Dimethylheptane	1
3,4 - Dimethylheptane	1
3,6 - Dimethylheptane	1
2,2 - Dimethylheptane	1
3,3 - Dimethylheptane	1
4,4 - Dimethylheptane	1
2,3,5 - Trimethylhexane	2
2,2,4 - Trimethylhexane	1
2,2,5 - Trimethylhexane	1
2,3,3 - Trimethylhexane	1
2,2,3,4 - Tetramethylpentane	2
2,3,3,4 - Tetramethylpentane	1
2,2,3,3 - Tetramethylpentane	9
3 - Ethylheptane	1
2,4 - Dimethyl - 3 - ethylpentane	1
3,3 - Diethylpentane	2
2,2,4,4 - Tetramethylpentane	1
	Ranking of correct structure out of 39 possible C_6NH_{15} isomers
Hexylamine	1
1,3 - Dimethylbutylamine	1
1,2,2 - Trimethylpropylamine	1
2,2 - Dimethylbutylamine	2
Dipropylamine	2
Diisopropylamine	1
N - Ethylbutylamine	1
N - Ethyl - sec - butylamine	1
Triethylamine	1
N,N - Dimethyl - sec - butylamine	6
N,N - Dimethyl - tert - butylamine	1